PAPERS ON CLIMATOLOGY IN RELATION TO AGRICULTURE, TRANSPORTATION, WATER RESOURCES, ETC.

CORRELATION.

By J. WARREN SMITH, Professor of Meteorology.

An attempt will be made in the following paper to explain some of the simplest methods of correlation. The relation of the weather in June and July to the yield of potatoes, in Ohio, will be used for a concrete example.

Figure No. 1 illustrates the common method of showing the correlation between two or more factors. A normal horizontal line is drawn through the center of the chart, figures at the left indicate the values above or below the normal and the years are given at the top of the page.

Dots are made on the line under each year opposite the proper departure figures and then these dots are connected by straight lines.

In this chart the solid line, A, shows the departure of the potato yield from the normal from year to year, in Ohio. The broken line, B, indicates the departure of the rainfall of Ohio for June and July of each year from the normal. The dotted line, C, shows the departure of the average temperature from the normal for the same months.

It is customary to say that when two curves run together—that is, when one runs above the normal line and the other does the same, and when one goes below the normal line and the other usually follows—they show a correlation between the factors.

On this chart, therefore, inasmuch as the lines A and B, representing, respectively, potato yield and rainfall, appear to follow each other in a general way, we may assume that there is some relation between the rainfall of June and July and the yield of potatoes.

On the other hand, a careful inspection of curves A and C indicates an opposite bending of the lines, and hence there must be a negative correlation.

The trouble with this method of correlation, however, is that while with two curves only a very general comparison can be made, with three or more curves there seems to be only a confusion of lines.

Figure No. 2: If three curves are to be used, it is far better to arrange the years so that one of the factors will have an increasing or decreasing value, as in diagram 2.

In this chart the years are arranged so that the curve showing the yield of potatoes runs from the highest yield regularly to the lowest. Then by drawing the other curves for the years as shown at the top broad general correlations can be shown. It will be found better in practice, however, to put only two curves on the same diagram.

Examination of the curves on this chart indicates a slight general relation between the yield and rainfall, and a strong opposite relation between the yield and temperature. When the yield is above the normal the temperature is nearly always below the normal, and when the yield is down the temperature is generally high.

The correlation between the temperature and yield is brought out much more clearly, however, by means of the dotted chart, as on figure 3.

On this the normal temperature is indicated by the horizontal line and the normal yield by the perpendicular line. At the right are figures showing temperature values above or below the normal and at the top figures indicating yields above or below the normal.

By taking the temperature and yield departures for each year and placing a dot on the chart where the lines indicating the departure values intersect, the grouping of the dots will show the correlation of the two factors.

If there is no close correlation between the two factors the dots will be scattered indiscriminately over the chart. If there is a relation, however, the dots will be grouped systematically, as shown on chart 3. It will be seen at a glance that the above-normal yield dots, with two exceptions, are grouped in the below-normal temperature side and the below-normal yield dots are, with one exception, all grouped on the above-normal temperature side.

Figure No. 4: In a study of the effect of weather conditions upon the growth of crops it is often necessary to combine the influence of two different factors, as rainfall and temperature. This can not be done by the curve method, but it can by a modification of the dotted chart.

In figure No. 4 the horizontal line and the figures at the right indicate temperature values, as in figure No. 3. The perpendicular line shows the average rainfall and the figures at the top the departure from the normal of the rain for June and July.

The upper left-hand quadrant shows conditions warm and dry; the upper right-hand quadrant, warm and wet; the lower right hand, cool and wet; the lower left, cool and dry.

By placing the yield dot at the intersection of the lines showing the departure of the rainfall and temperature from the normal, for each year, we have a chart giving the combined effect of rainfall and temperature upon the yield.

If we indicate the yields above normal by crosses and the yields below normal by the minus signs, a very decided grouping will be found if there is a distinct relation between the factors. In practice it will be found better to use red and blue dots to indicate yields above or below the normal and to place the yield-departure figures beside each dot.

By this figure it can be seen at a glance that there has never been a poor potato yield in Ohio in the 27 years under discussion when June and July have averaged cool and wet, and but one poor yield when these months have been cool and dry.

Also that there has been but one year when these months have averaged warm and wet, and that year the crop was poor. It takes but a glance also to see that warm and dry weather in June and July is distinctly unfavorable for potatoes.

While both the curve and the dotted diagrams are of value to illustrate the relation that has been determined and are of value oftentimes in finding whether there may be a correlation between two factors, neither of them answers the question as to the exact value of the correlation.

In figure No. 4, for example, it seems perfectly plain that there is an inverse relation between the temperature and the potato yield, but it is impossible to say positively whether the rainfall is favorable or not. There are nine crosses on one side of the normal precipitation line and eight on the other.

To ascertain this and answer questions of correlation in general, the writer has made use of the common correlation table as shown in Tables 1 and 2.

It is not the province of this paper to go into a mathematical discussion of these tables because the principle on which they operate is well established. It is sufficient to say that if the correlation coefficient, r, as determined by the table is $+1$ there is an exact correlation between the factors, and if the correlation coefficient is -1 there is an exact negative or opposite relation.

If, on calculation, r is found to be near zero, then there is little or no correlation between the two factors. In practice it is safe to say without question that there is a well-defined relation if the correlation coefficient is six times the probable error.

The probable error is found by the following equation, in which r is the correlation coefficient and n is the number of years under consideration:

$$0.674\frac{1-r}{\sqrt{n}}$$

In Table 1 the correlation coefficient is found to be 0.50, which shows a fairly well-defined relation between the rainfall for June and July and the potato yield. This is not quite six times the probable error, which is ± 0.097, yet it shows the beneficial effect of the rainfall much more conclusively than either the curve or the dotted diagrams.

In Table 2 the value of r is found to be -0.69 and the probable error ± 0.10. This indicates a very strong inverse relation between the temperature for June and July and the yield of potatoes, as would be expected from the curve and dotted diagrams.

From these charts and tables the conclusion is that June and July should be moderately wet and must be cool to produce a good potato yield in Ohio.

EXPLANATION OF FIGURE NO. 3.

The horizontal line indicates the average temperature for Ohio for June and July. The figures at the right show temperature values above or below the normal. The perpendicular line is drawn for the average yield of potatoes, and the figures at the top of the chart indicate yields above or below the normal in bushels per acre. To make this chart, place a dot at the intersection of the lines showing the departure of the temperature and of the yield from the normal. If there is no connection between the weather element and the yield the dots will be scattered all over the diagram. In this chart practically all of the yields above the normal fall into the quadrant showing temperatures below normal, while the yields below normal correspond with temperatures above normal.

EXPLANATION OF FIGURE NO. 4.

This is used for correlating three factors. The average temperature for June and July is shown by the horizontal line and the values above or below the normal by the figures at the right. The perpendicular line shows the average rainfall for June and July, with the figures on each side indicating the amounts above or below the normal. To work this chart, place a dot for each year where the temperature and rainfall value lines intersect, marking a plus sign if the yield is above the normal and a minus sign if below normal. In practice it will be found better to indicate the above-normal yield values by red dots and the below normal in blue. It is of advantage also to put the figures showing the amount of the yield above or below the normal beside each dot. This chart shows the combined effect of rainfall and temperature upon the crop yield.

TABLE 1.—*Rainfall and potato yield, Ohio, 1883 to 1909.*

Year.	Rainfall.			Yield.			Product of columns 3 and 6.
	Rainfall.	Departure.	Square of column 3.	Yield.	Departure.	Square of column 6.	
1	2	3	4	5	6	7	8
	Inches.	*Inches.*		*Bushels.*	*Bushels.*		
1883	8.4	+0.3	0.09	99	+25	625	+ 7.5
1884	6.8	−1.3	1.69	75	+ 1	1	− 1.3
1885	7.5	− .6	.36	75	+ 1	1	− .6
1886	6.4	−1.7	2.89	78	+ 4	16	− 6.8
1887	5.9	−2.2	4.84	30	−44	1,936	+ 96.8
1888	7.8	− .3	.09	80	+ 6	36	− 1.8
1889	8.3	+ .2	.04	75	+ 1	1	+ .2
1890	6.5	−1.6	2.56	46	−28	784	+ 44.8
1891	8.6	+ .5	.25	98	+24	576	+ 12.0
1892	9.4	+1.3	1.69	60	−14	196	− 18.2
1893	5.8	−2.3	5.29	58	−16	256	+ 36.8
1894	4.2	−3.9	15.21	63	−11	121	+ 42.9
1895	4.5	−3.6	12.96	63	−11	121	+ 39.6
1896	12.9	+4.8	23.04	89	+15	225	+ 72.0
1897	7.5	− .6	.36	42	−32	1,024	+ 19.2
1898	6.8	−1.3	1.69	61	−13	169	+ 16.9
1899	7.1	−1.0	1.00	71	− 3	9	+ 3.0
1900	7.6	− .5	.25	76	+ 2	4	− 1.0
1901	7.2	− .9	.81	54	−20	400	+ 18.0
1902	12.2	+4.1	16.81	94	+20	400	+ 82.0
1903	7.7	− .4	.16	83	+ 9	81	− 3.6
1904	7.0	−1.1	1.21	98	+24	576	− 26.4
1905	8.6	+ .5	.25	78	+ 4	16	+ 2.0
1906	8.6	+ .5	.25	110	+36	1,296	+ 18.0
1907	9.9	+1.8	3.24	76	+ 2	4	+ 3.6
1908	6.6	−1.5	2.25	77	+ 3	9	− 4.5
1909	9.6	+1.5	2.25	96	+22	484	+ 33.0
Summary			101.53			9,367	+484.1

EXPLANATION OF TABLE.

Column No. 1 shows the years; No. 2 the total rainfall for Ohio for June and July, in inches and tenths; No. 3 the departure from the normal rainfall for these two months, the normal for 54 years being 8.1 inches; No. 4 the square of the departure as shown in column 3; No. 5 gives the yield of potatoes for Ohio in bushels per acre; No. 6 the departure from the normal potato yield, the average potato yield for the 27 years under consideration being 74 bushels per acre; No. 7, the square of the departure in column 6; column No. 8 shows the product of the departures in columns 3 and 6. Care must be taken to precede the product by the proper sign in this column, merely remembering that the product of two numbers with like signs is preceded by a plus sign, and the product of two numbers with unlike signs with a minus sign.

To determine the correlation coefficient, r, divide the algebraic sum of the values in column 8 by the square

root of the product of the sums of columns 4 and 7. The calculation in this case is $101.53 \times 9{,}367 = 951{,}031.51$. The square root of this product is 975.20. 484.1 divided by $975.20 = 0.50$, the correlation coefficient. This is moderately high, although it is not quite six times the probable error, which is ± 0.097.

TABLE 2.—*Temperature and potato yield, Ohio, 1883 to 1909.*

Year.	Temperature.			Yield.			Product of columns 3 and 6.
	Temperature.	Departure.	Square of column 3.	Yield.	Departure.	Square of column 6.	
1	2	3	4	5	6	7	8
1883	70.6	−0.8	0.64	99	+25	625	− 20.0
1884	71.3	− .1	.01	75	+ 1	1	− .1
1885	71.2	− .2	.04	75	+ 1	1	− .2
1886	69.8	−1.6	2.56	78	+ 4	16	− 6.4
1887	74.4	+3.0	9.00	30	−44	1,936	−132.0
1888	71.2	− .2	.04	80	+ 0	36	− 1.2
1889	69.6	−1.8	3.24	75	+ 1	1	− 1.8
1890	73.2	+1.8	3.24	46	−28	784	− 50.4
1891	70.0	−1.4	1.96	98	+24	576	− 33.6
1892	73.0	+1.6	2.56	60	−14	196	− 22.4
1893	72.6	+1.2	1.44	58	−16	256	− 19.2
1894	72.8	+1.4	1.96	63	−11	121	− 15.4
1895	71.8	+ .4	.16	63	−11	121	− 4.4
1896	71.4	.0	0	89	+15	225	.0
1897	71.8	+ .4	.16	42	−32	1,024	− 12.8
1898	71.0	− .4	.16	61	−13	169	+ 5.2
1899	72.8	+1.4	1.96	71	− 3	9	− 4.2
1900	72.0	+ .6	.36	76	+ 2	4	+ 1.2
1901	74.5	+3.1	9.61	54	−20	400	− 62.0
1902	70.4	−1.0	1.00	94	+20	400	− 20.0
1903	68.6	−2.8	7.84	83	+ 9	81	− 25.2
1904	69.9	−1.5	2.25	98	+24	576	− 36.0
1905	71.1	− .3	.09	78	+ 4	16	− 1.2
1906	71.0	− .4	.16	110	+36	1,296	− 14.4
1907	69.1	−2.3	5.29	76	+ 2	4	− 4.6
1908	71.6	+ .2	.04	77	+ 3	9	+ .6
1909	70.4	−1.0	1.00	96	+22	848	− 22.0
Summary			56.77			9,367	−502.5

TABLE 2—*Temperature and potato yield, Ohio, 1883 to 1909*—Contd.

EXPLANATION OF TABLE.

This table is similar to table No. 1 except that columns 2 to 4 contain the average temperature data for Ohio for the months of June and July, instead of the rainfall as in Table 1. To determine the correlation coefficient multiply 56.77 by 9,367 = 531,764.59. The square root of this is 729.22 − 502.5 divided by 729.22 = −0.69, quite a high negative coefficient. This is safely above six times the probable error, which in this case is ± 0.10. This indicates that the temperature for June and July has quite a high negative influence on the potato yield.

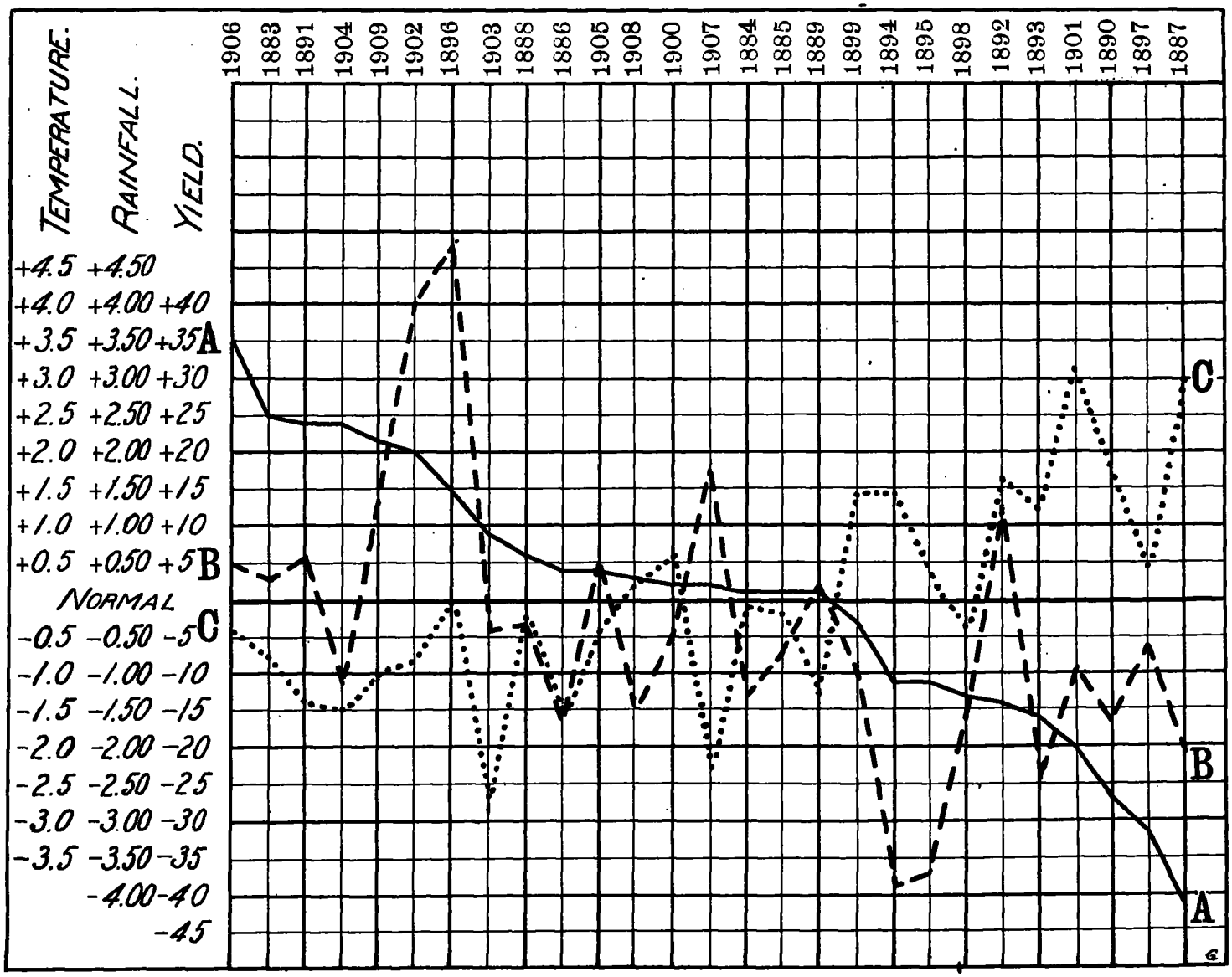

FIG. 1.—Relation of weather to yield of potatoes in Ohio, 1883–1909.
A—Yield of potatoes in bushels per acre.
B—Departure of rainfall from normal for June and July.
C—Departure of temperature from the normal for June and July.

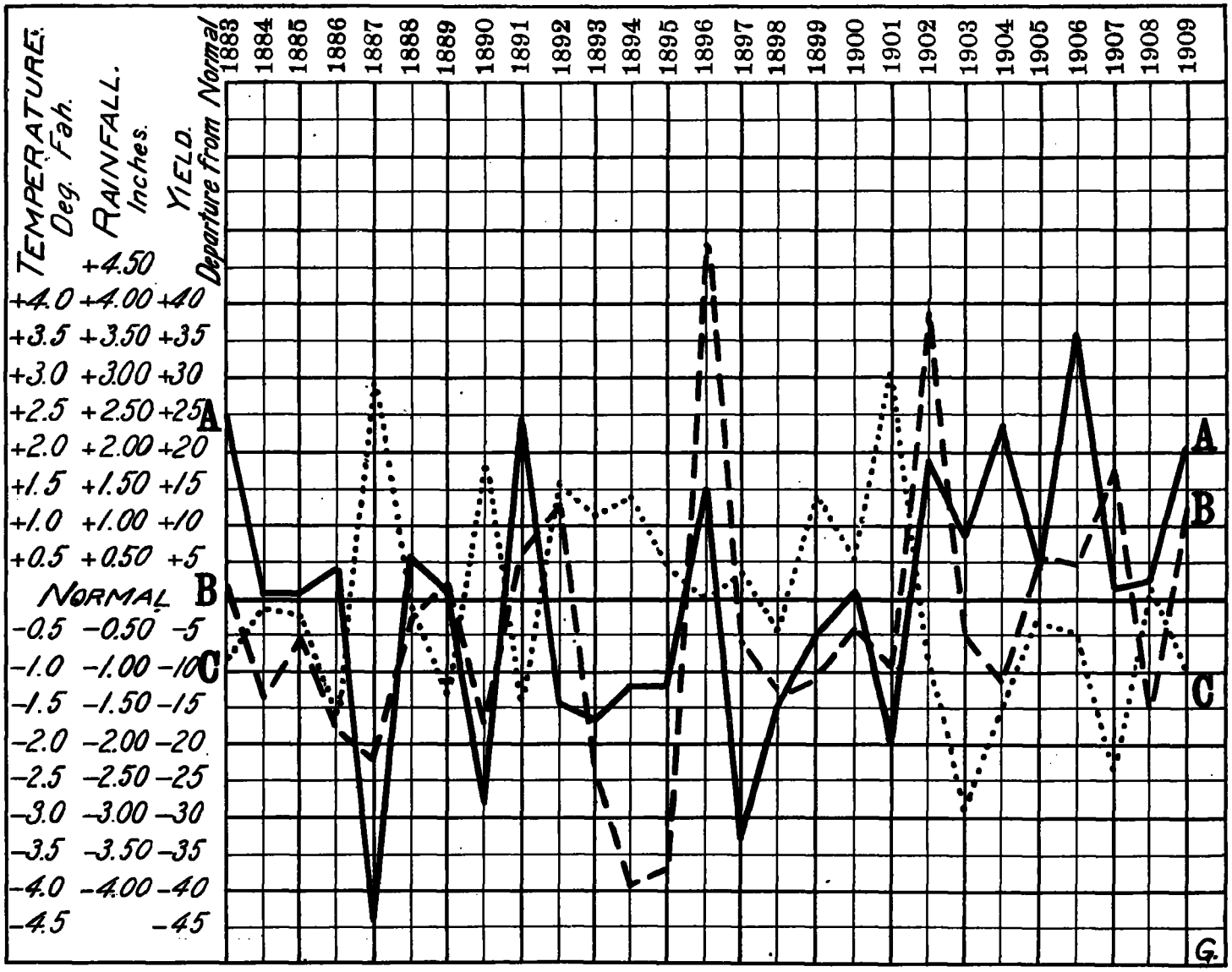

FIG. 2.—Relation of weather to yield of potatoes in Ohio 1883–1909.
A—Yield of potatoes in bushels per acre, departure from normal.
B—Rainfall departure from normal for June and July.
C—Temperature departure from normal for June and July.

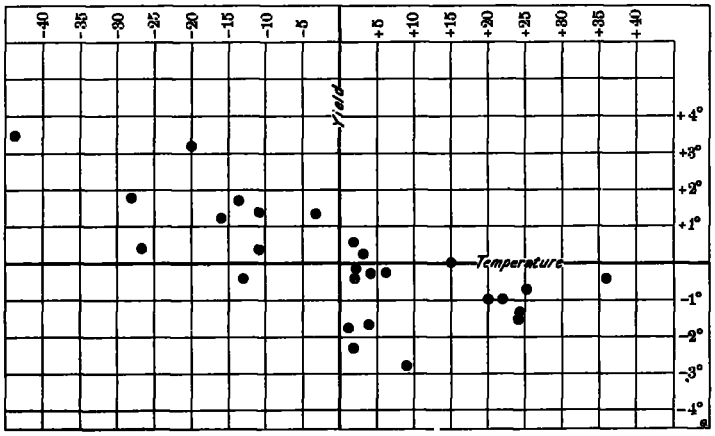

FIG. 3.—Temperature and yield of potatoes in Ohio, 1883–1909.

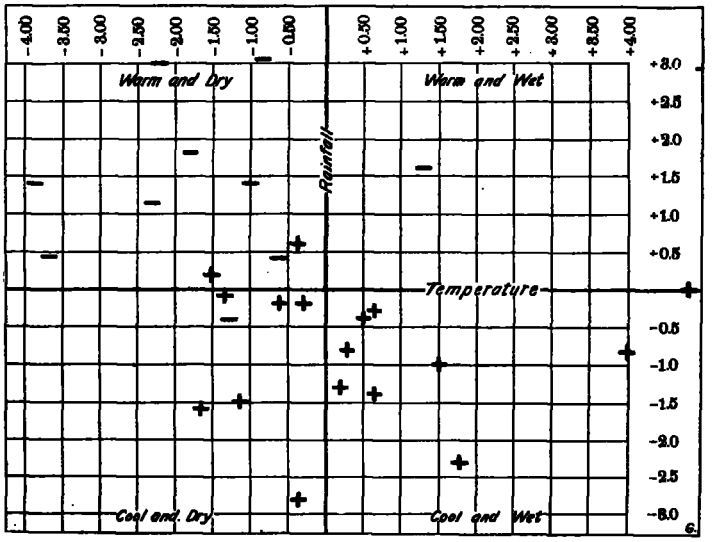

FIG. 4.—Relation of weather to yield of potatoes in Ohio, 1883–1909.